

Introduction to Data Analysis with Orange

Amy Larner Giroux, PhD

UCF Center for Humanities & Digital Research

NEH Digital Culture Summer Institute



Welcome

The tutorial outlined in this slide deck will step you through your first workflow in Orange. This is the short and sweet version. For the full details, please consult the “Orange Tweet Analysis Tutorial” PDF available in Google Classroom. That document covers both tutorials for this week in minute detail.

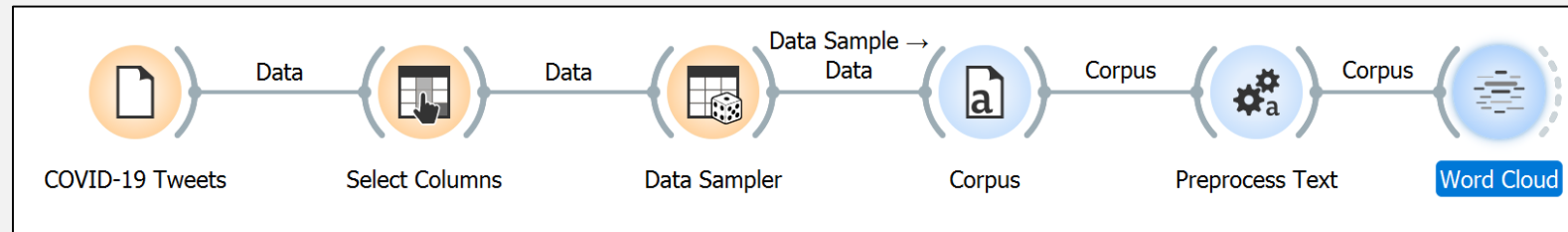
Agenda

- Brief slice of Orange
- Discussion on categorizing your data (this is for discussion only, not to be done on your data this week)
- General tips and concepts in Orange
- Detailed steps for your first workflow – from loading your data file to producing a word cloud
- Your NEH Institute deliverables for this tutorial



Orange

- Data mining toolset (text, images, networks, etc.)
- Workflow process
- Widget-based
- No programming is necessary
- You create workflows by connecting widgets



If you have not already installed Orange and the Text Add-on, please see the detailed directions in the PDF.

Pre-Processing Your Data

- Data preparation will be 80-90% of your workflow leading up to using Orange for analysis. The old adage of “garbage in equals garbage out” pertains to any project using data and tweets are no exception.
- For you to be able to interpret the results of any type of textual analysis of your data, you need to become intimately familiar with the content, regardless of the size of the dataset.
- This does not mean that you necessarily have to read every tweet in a 10,000+ tweet dataset, but you should sample the data for a set to read closely and choose ways to categorize your data for more comparative analyses.
- This categorization of your data will help you to answer your research questions.

You are not expected to complete this level of data preparation on your own data for this tutorial. Read through the information to understand it in terms of the sample dataset but do not try to perform these steps with your own data until after the Institute is over.



Sample Dataset

The sample screens in this tutorial are from a dataset that we will use in the second tutorial.

Throughout the current tutorial, please use the tweet data sent to you prior to the Institute.

	A	B	C	D	E	F	G	H
1	Date	Text	User	To	Retweets	Favorites	Mentions	Hashtags
2	3/23/2020	.@realDonaldTrump just said: "if it were up to the	DrLeanaWen		24641	100067	@realDonaldTrump	#covid19
3	3/16/2020	There it is. I've been deathly afraid of this exact	eugenegu	realDon	24532	216814		
4	3/19/2020	America's Private Sector is stepping up to help us	realDonaldTrump		20931	86850		#COVID19
5	3/12/2020	WOW. Rep. Porter just read the testing costs to	girlsreallrule		18318	57843	@realDonaldTrump	#Covid_19
6	3/21/2020	For anyone who read the tweet from	tedlieu	EdselSal	17937	37171	@realDonaldTrump	#Covid_19

The dataset for the screen shots is from a project where we wanted to focus on groups with opposing political ideologies and used COVID19 and @realDonaldTrump as our search criteria.

Dataset contains 1,121 tweets from this criteria that were retweeted 100 or more times.



Categorizing Your Data

As you read through the reasoning for the categorization chosen for the sample dataset, think about how your tweet dataset could be categorized.

Sample dataset

- we looked at the specific users who sent these tweets and the dataset was comprised of 543 unique individuals
- Twitter profile data for these individuals was collected and matched with the tweets in the dataset
- general content of each user's tweets and number of followers were used to categorize the users in three ways:
 - Influencers (followers > 10,000)
 - Opinion-leaders (users who sent informational tweets including links to articles/videos)
 - Political leaning (left, neutral, right, based on profile information and tweet content)
- Overall categorization was assigned to each individual tweet
- Any given participant could be scored as both an influential and an opinion-leader.

When you are ready to work on your data after the Institute, please see the detailed directions in the PDF on how to use pivot tables and VLOOKUP in Excel to set up your data categorization.

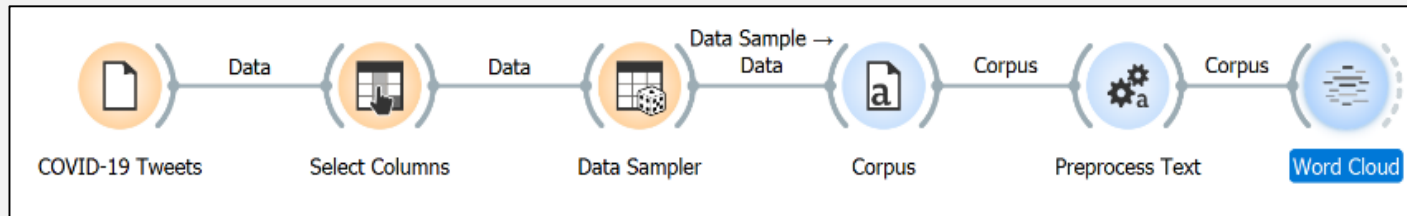


Orange Tips

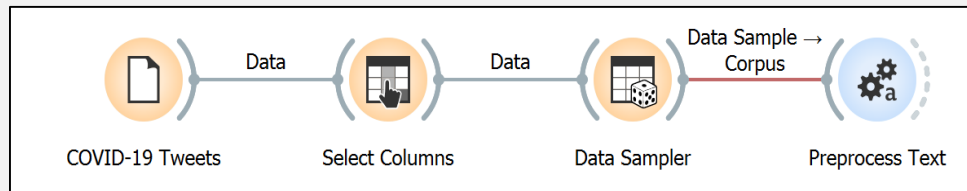
Like any piece of software, Orange has nuances in the process that can sometimes be confusing. The following list are some of the things to be aware of so that you recognize what is happening in the program.

Widget input-output data types

- The connecting lines between the widgets display the type of data passed between. In this example “data” is used between the first 4 widgets and then the Corpus widget converts the “data” to “corpus” and that is used in the last two widgets.



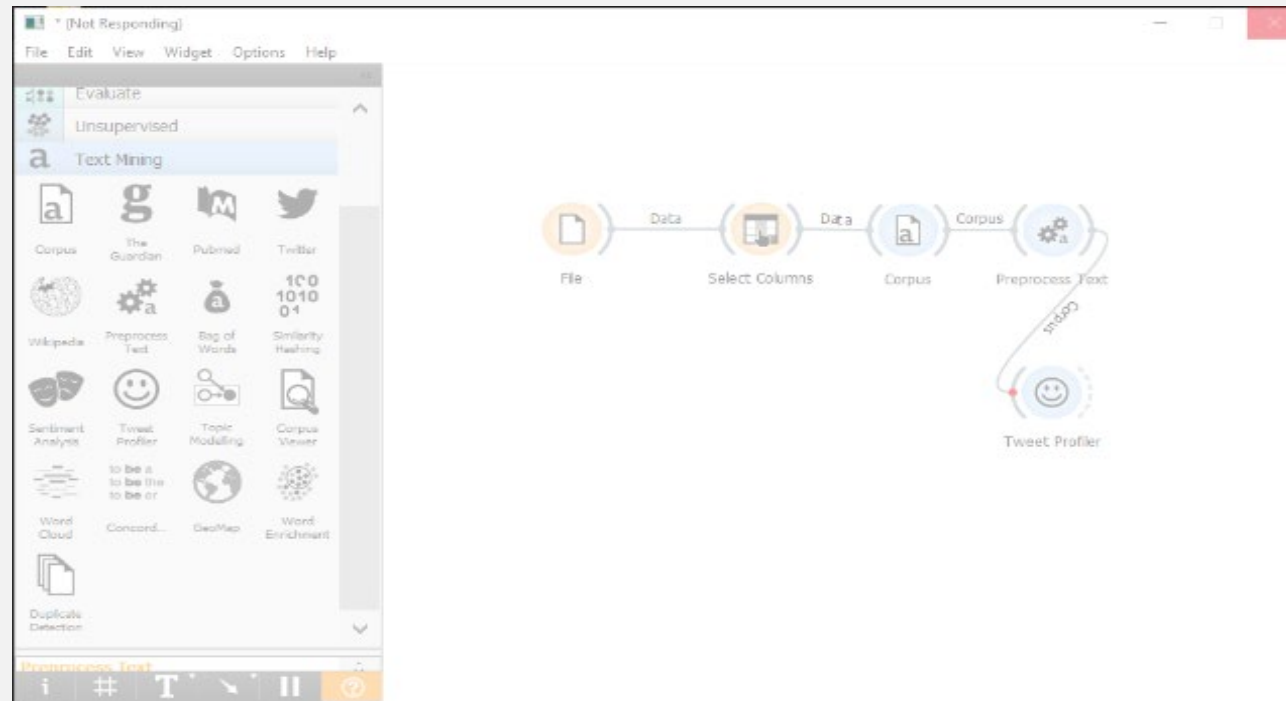
- If you try to connect widgets that do not have the same input requirements, the connecting line will turn red to notify you of the issue.



Be Patient

Depending on the size of the dataset, processing of the steps in an Orange workflow can take some time. The default for all the widgets is to process automatically when connected into the workflow.

At times this can cause the program to show (Not Responding) in the title bar and grey itself out as in this example. The red dot next to the Tweet Profiler widget shows that it is processing data. Be patient and let the process finish.

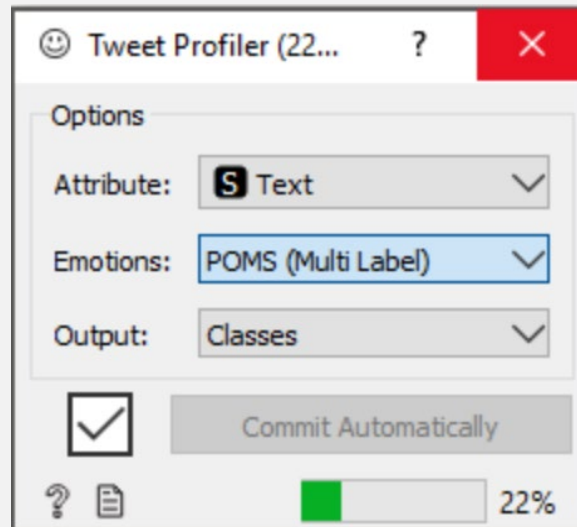


Taking Control

Most widgets have an option that you can uncheck to prevent them from running automatically.

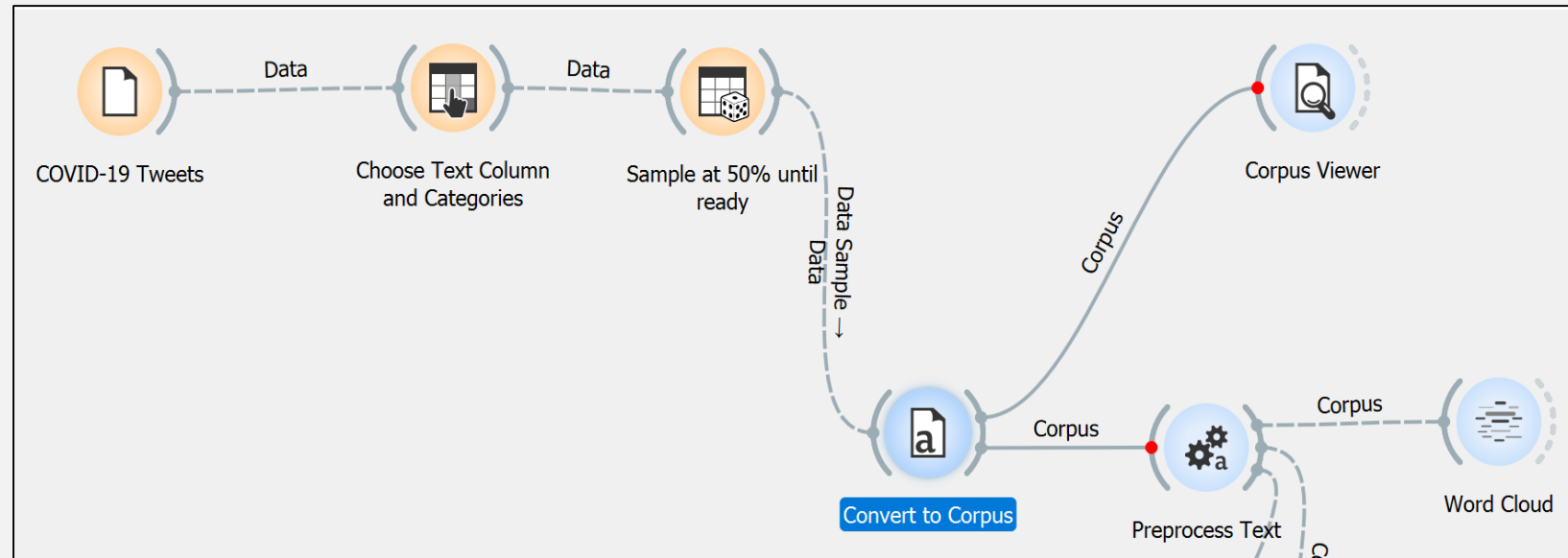
This example from the Tweet Profiler has a checkbox next to the “Commit Automatically” button. Unchecking the box allows you to run the widget when you want to by clicking the button.

The terminology of the buttons is inconsistent. For example, in Sentiment Analysis the checkbox makes the button have “Autocommit is on.” Just be aware of the concept and look for the checkbox/button combinations.



Renaming Widgets

You can rename the widgets to make notes to yourself about the process.

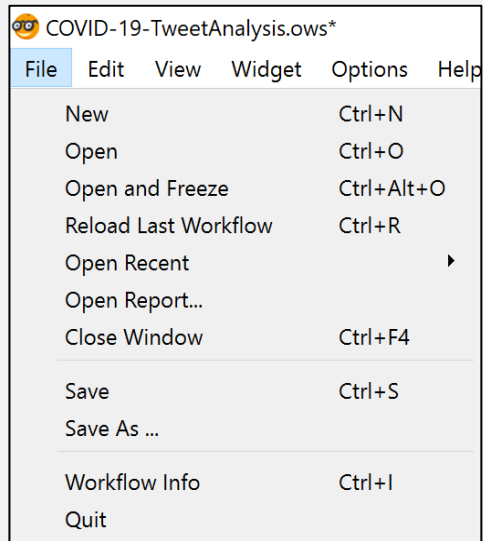


Either right-click on the widget and select Rename from the popup menu, or click once on the widget to select it and press F2.

Opening Workflows

There are three options to open a saved workflow.

1. **Double-click** on the .ows file in File Explorer
2. Open Orange and use **Ctrl-O** or **File -> Open**
3. Open Orange and use **Ctrl-Alt-O** or **File -> Open and Freeze**

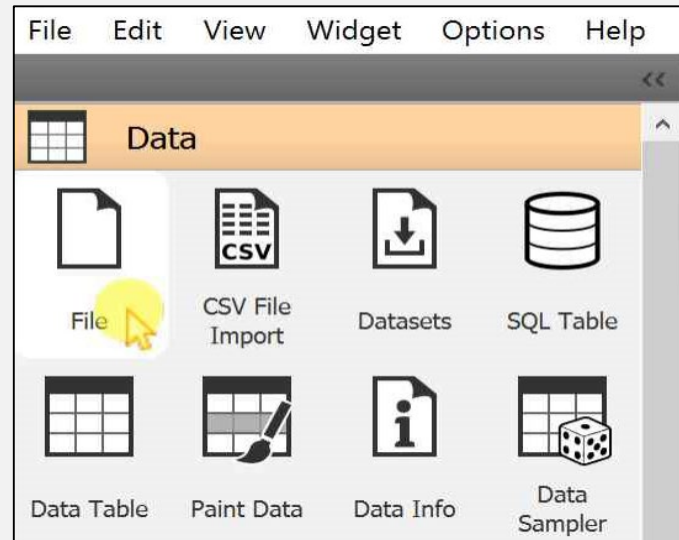


To prevent the automated workflow from running immediately, you can use the Open and Freeze.



Last Caveats

- There are not a lot of Finish or OK buttons on various screens within Orange. Sometimes you will leave a window open while you do other tasks and watch how those tasks affect your output (e.g. Word Cloud).
 - There will be comments in the tutorial text that will let you know when it is safe to close a window.
- There are multiple ways to add widgets to a workflow, such as double or right-clicking on the work area or selecting from the widget window.
 - This tutorial will use the widget window but you are welcome to use the other methods.



Let's Get Started!

For this hands-on portion, please use the tweet data that was emailed to you that contained the hashtags and date range you requested.

Open your CSV file in Excel:

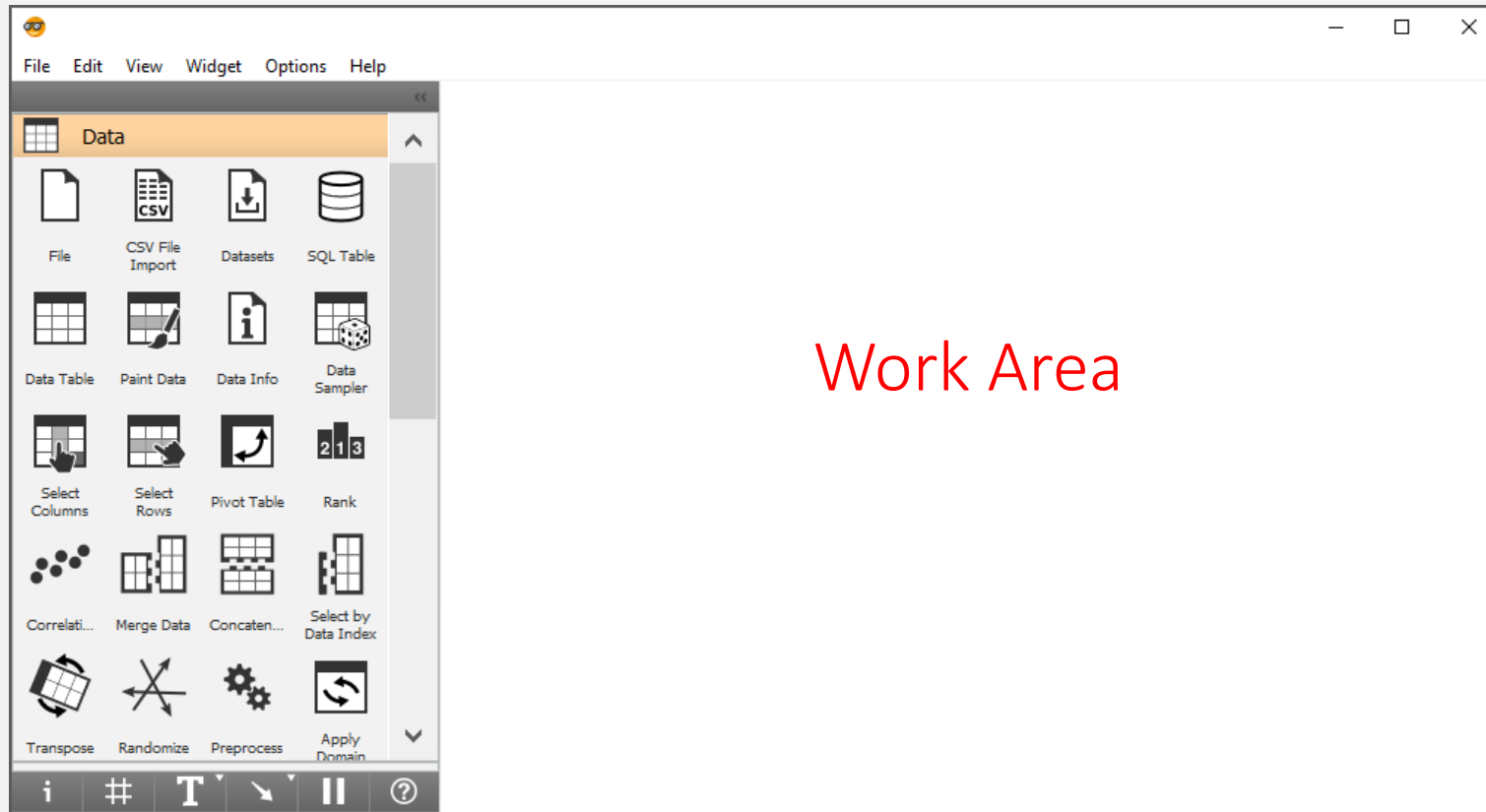
1. delete the first column (the count)
2. delete the second column (ID)
3. save the file as an Excel spreadsheet (.xlsx)

	A	B	C	D	E	F	G	H
1	Date	Text	User	To	Retweets	Favorites	Mentions	Hashtags
2	3/23/2020	.@realDonaldTrump just said: "if it were up to the	DrLeanaWen		24641	100067	@realDonaldTrump	#covid19
3	3/16/2020	There it is. I've been deathly afraid of this exact	eugenegu	realDonaldTrump	24532	216814		
4	3/19/2020	America's Private Sector is stepping up to help us	realDonaldTrump		20931	86850		#COVID19
5	3/12/2020	WOW. Rep. Porter just read the testing costs to	girlsreallyrule		18318	57843	@realDonaldTrump	#Covid_19
6	3/21/2020	For anyone who read the tweet from	tedlieu	EdselSal	17937	37171	@realDonaldTrump	#Covid_19

Your columns should look like this.



Open Orange



Widget toolbox

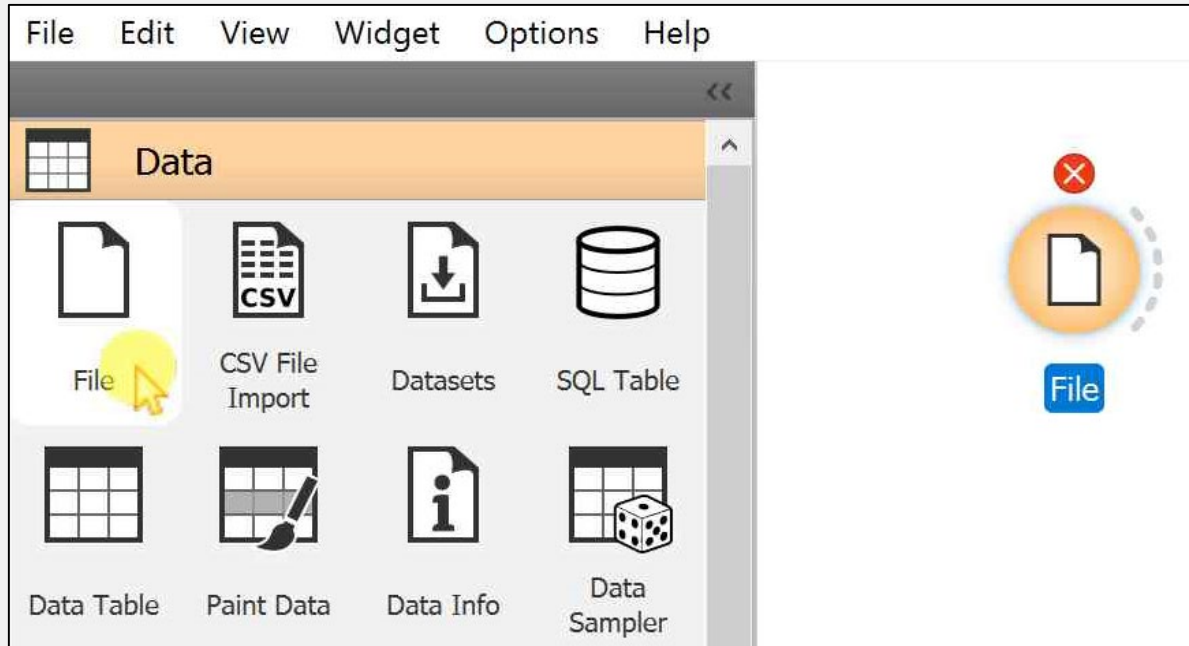
The widget toolbox is defaulted to be open. As you get familiar with Orange you can minimize it to have more room in the work area.



Use **Ctrl-S** or **File -> Save** to save your workflow. Do this fairly often to ensure you don't lose your work.

File Load

Click on **File** in the Data section of the widget window to add a File widget to the work area.

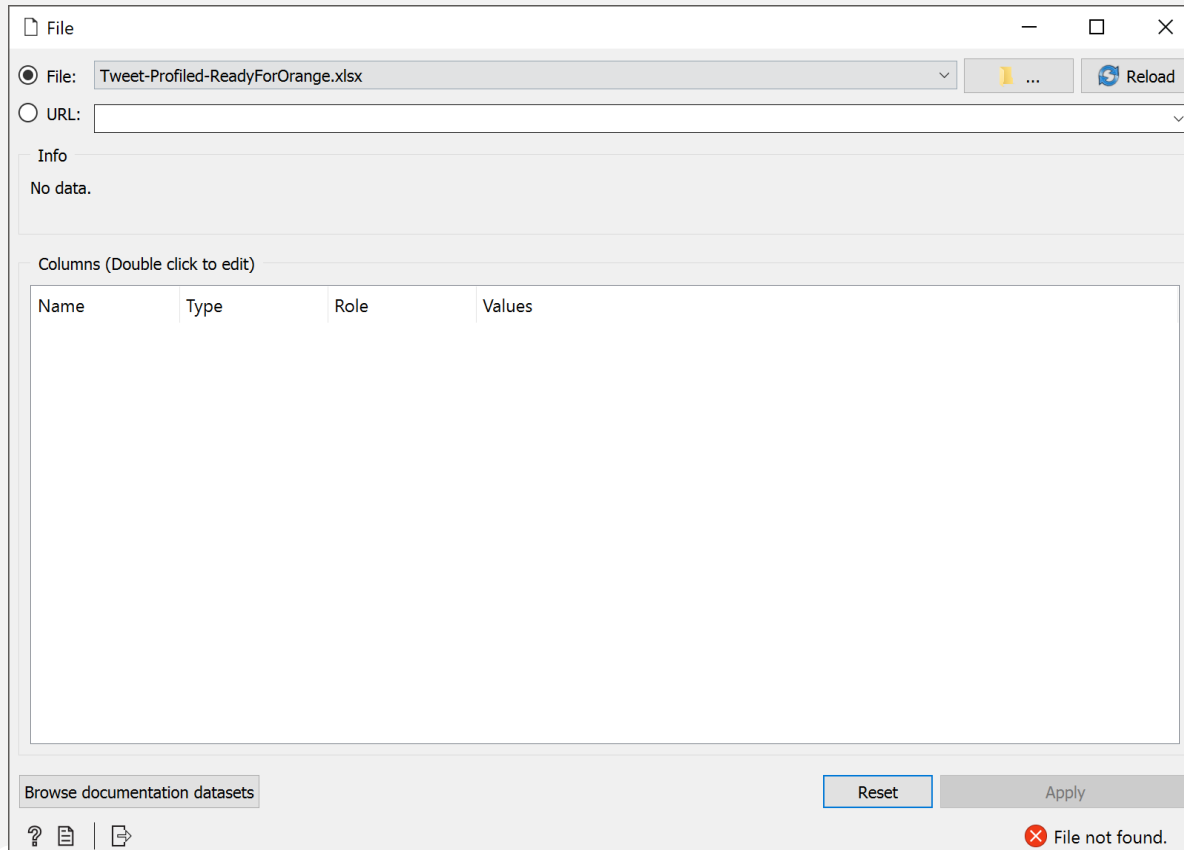


The red X just means a file hasn't been loaded yet.



File Load

Double-click on the File widget to open it.

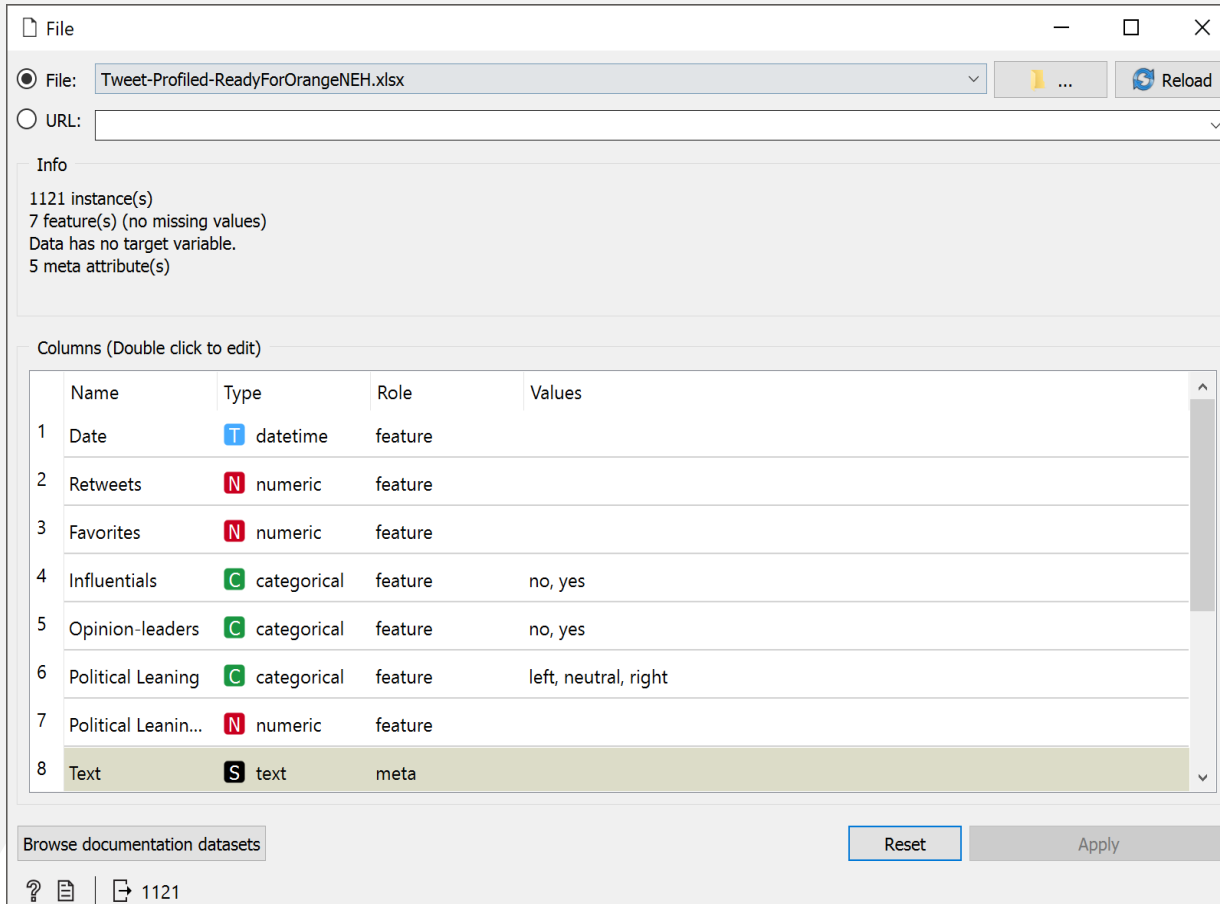


If you have opened files in other workflows, the last file may be shown. In the bottom right of this window it will show whether that file still exists. In this example the Tweet-Profiled-ReadyForOrange spreadsheet was not found.



File Load

Using the **open folder button**, browse and choose your spreadsheet



The screenshot shows the 'File' window in Orange Data Mining. The 'File' tab is selected, showing the file 'Tweet-Profiled-ReadyForOrangeNEH.xlsx'. The 'Info' section displays: 1121 instance(s), 7 feature(s) (no missing values), Data has no target variable, and 5 meta attribute(s). The 'Columns (Double click to edit)' table lists the following columns:

	Name	Type	Role	Values
1	Date	datetime	feature	
2	Retweets	numeric	feature	
3	Favorites	numeric	feature	
4	Influentials	categorical	feature	no, yes
5	Opinion-leaders	categorical	feature	no, yes
6	Political Leaning	categorical	feature	left, neutral, right
7	Political Leanin...	numeric	feature	
8	Text	text	meta	

At the bottom, there is a 'Browse documentation datasets' button, 'Reset' and 'Apply' buttons, and a status bar showing a question mark, document icon, and '1121'.

After opening your data file and examining the column information, you can close this window.

The columns of the spreadsheet will be displayed with the data type, role, and values determined by Orange. In some instances Orange chooses a different type than you want and you can double-click on the type column to change it.

Additionally on this screen it will show you the number of rows (1121 instances), the number of fields it deems are features (e.g. numeric or categorical) and the number of text fields (meta).

If you change data in the underlying spreadsheet, use the Reload button to reimport the data.

